Collaborative Tagging and Semiotic Dynamics

Ciro Cattuto*

Museo Storico della Fisica e Centro Studi e Ricerche "Enrico Fermi"
Compendio Viminale, 00184 Roma, Italy and
Dipartimento di Fisica, Università di Roma "La Sapienza"
P.le A. Moro, 2, 00185 Roma, Italy

Vittorio Loreto and Luciano Pietronero

Dipartimento di Fisica, Università di Roma "La Sapienza"
P.le A. Moro, 2, 00185 Roma, Italy
(Dated: February 1, 2008)

Collaborative tagging has been quickly gaining ground because of its ability to recruit the activity of web users into effectively organizing and sharing vast amounts of information. Here we collect data from a popular system and investigate the statistical properties of tag co-occurrence. We introduce a stochastic model of user behavior embodying two main aspects of collaborative tagging: (i) a frequency-bias mechanism related to the idea that users are exposed to each other's tagging activity; (ii) a notion of memory – or aging of resources – in the form of a heavy-tailed access to the past state of the system. Remarkably, our simple modeling is able to account quantitatively for the observed experimental features, with a surprisingly high accuracy. This points in the direction of a universal behavior of users, who – despite the complexity of their own cognitive processes and the uncoordinated and selfish nature of their tagging activity – appear to follow simple activity patterns.
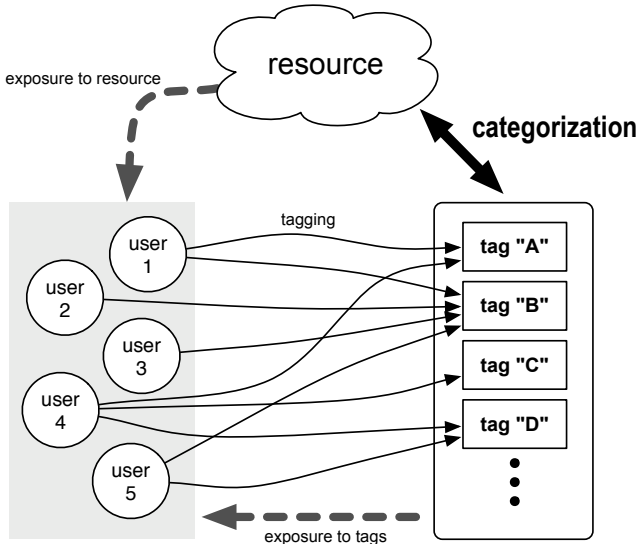
FIG. 1: Schematic depiction of the collaborative tagging process: web users are exposed to a resource and freely associate tags with it. Their interaction with the system also exposes them to tags previously entered by themselves and by other users. The aggregated activity of users leads to an emergent categorization of resources in terms of tags shared by a community.

1. INTRODUCTION

Recently, a new paradigm has been quickly gaining ground on the World-Wide Web: Collaborative Tagging [1, 2, 3]. In web applications like *del.icio.us*[1], *Flickr*[2], *CiteULike*[3], *Connotea*[4], users manage, share and browse collections of online resources by enriching them with semantically meaningful information in the form of freely chosen text labels (*tags*). The paradigm of collaborative tagging has been successfully deployed in web applications designed to organize and share diverse online resources such as bookmarks, digital photographs, academic papers, music and more. Web users interact with a collaborative tagging system by posting content (*resources*) into the system, and associating text strings (*tags*) with that content, as shown in Fig. 1. At the global level the set of tags, though determined with no explicit coordination, evolves in time and leads towards patterns of terminology usage that are shared by the entire user community. Hence one observes the emergence of a loose categorization system – commonly referred to as *folksonomy* – that can be effectively used to navigate through a large and heterogeneous body of resources.

Focusing on tags as basic dynamical entities, the process of collaborative tagging falls within the scope of Semiotic Dynamics [4, 5], a new field that studies how populations of humans or agents can establish and share semiotic systems, typically driven by their use in com-

*Electronic address: ciro.cattuto@roma1.infn.it

munication. Indeed the emergence of a folksonomy exhibits dynamical aspects also observed in human languages [6, 7], such as the crystallization of naming conventions, competition between terms, takeovers by neologisms, and more.

In the following we adopt the point of view of complex systems science and try to understand how the "microscopic" tagging activity of users causes the emergence of the high-level features we observe for the ensuing folksonomy. We ground our analysis on actual tagging data extracted from *del.icio.us* and *Connotea* and use standard statistical tools to gain insights into the underlying tagging dynamics. Based on this, we introduce a simple stochastic model for the tagging behavior of an "average" user, and show that such a model – despite its simplicity – is able to reproduce extremely well some of the observed properties. We close giving an interpretation of the model parameters and pointing out directions for future research.

2. EXPERIMENTAL DATA

The activity of users interacting with a collaborative tagging system consists of either navigating the existing body of resources by using tags, or adding new resources to the system. In order to add a new resource to the system, the user is prompted for a reference to the resource and a set of tags to associate with it. Thus the basic unit of information in a collaborative tagging system is a (`user, resource, {tags}`) triple, here referred to as *post*. Tagging events build a tri-partite graph (with partitions corresponding to users, resources and tags, respectively) and such a graph can be subsequently used as a navigation aid in browsing tagged information. Usually a post contains also a temporal marker indicating the physical time of the tagging event, so that temporal ordering can be preserved in storing and retrieving posts.

Our analysis will focus on *Del.icio.us*, for several reasons: i) it was the very first system to deploy the ideas and technologies of collaborative tagging, and the paradigmatic character it acquired makes it a natural starting point for any quantitative study. ii) because of its popularity, it has a large community of active users and comprises a precious body of raw data on the static and dynamical properties of a folksonomy. iii) it is a *broad folksonomy* [8], and single tagging events (posts) retain their identity and can be individually retrieved. This affords unimpeded access to the "microscopic" dynamics of collaborative tagging, providing the opportunity to make contact between emergent behaviors and low-level dynamics. It also allows to define and measure the multiplicity (or frequency) of tags in the context of a single resource. Contrary to this, popular sites falling in the *narrow folksonomy* class (*Flickr*, for example) foster a different model of user interaction, where tags are mostly applied by the content creator, no notion of tag multiplicity is available in the context of a resource, and

TABLE I: Statistics of the datasets used for the co-occurrence analysis. For each tag in the first column we report the number of posts marked with that tag, the number of total and distinct tags co-occurring with it, and the corresponding number of resources. The data were retrieved during May 2005.

Tag	No. posts	No. tags	No. distinct tags	No. resources
Blog	37974	124171	10617	16990
Ajax	33140	108181	4141	2995
Xml	24249	108013	6035	7364
H5N1	981	5185	241	969

no access is given to the raw sequence of tagging events.

On studying *Del.icio.us* we adopt a tag-centric view of the system, that is we investigate the evolving relationship between a given tag and the set of tags that co-occur with it. In line with our focus on semiotic dynamics, we factor out the detailed identity of the users involved in the process, and only deal with streams of tagging events and their statistical properties. To perform automated data collection of raw data we use a custom web (HTTP) client that connects to *del.icio.us* and navigates the system's interface as an ordinary user would do, extracting the relevant metadata and storing it for further post-processing. *del.icio.us* allows the user to browse its content by tag: our client requests the web page associated with the tag under study and uses an HTML parser to extract the post information (user, resource, tags, time stamp) from the returned HTML code. Fig. 2 graphically depicts the raw data we gather, for the case of two popular tags on *del.icio.us*. Table I describes the datasets we used for the present analysis.

3. DATA ANALYSIS

Here we analyze data from *del.icio.us* and *Connotea* and investigate the statistical properties of tag association. Specifically, we select a semantic context by extracting the resources associated with a given tag X and study the statistical distribution of tags co-occurring with X (see Table 1). Fig. 2 graphically illustrates the associations between tags and posts, and Fig. 3 reports the frequency-rank distributions for the tags co-occurring with a few selected ones. The high-rank tail of the experimental curves displays a power-law behavior, signature of an emergent hierarchical structure, corresponding to a generalized Zipf's law [9] with an exponent between 1 and 2. Since power laws are the standard signature of self-organization and of human activity [10, 11], the presence of a power-law tail is not surprising. The observed value of the exponent, however, deserves further investigation because the mechanisms usually invoked to explain Zipf's law and its generalizations [12] don't look very realistic for the case at hand, and a mechanism grounded on experimental data should be sought.
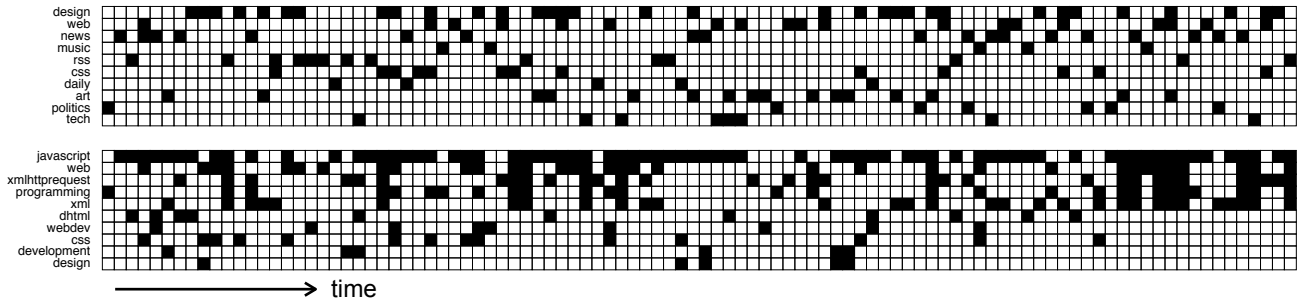
FIG. 2: Tagging activity: a time-ordered sequence of tagging events is graphically rendered by marking the tags co-occurring with *blog* (top panel) or *ajax* (bottom panel) in an experimental sequence of posts on *del.icio.us*. In each panel, columns represent single tagging events (posts) and rows correspond to the 10 most frequent tags co-occurring with either *blog* (top panel) or *ajax* (bottom panel). 100 tagging events are shown in each panel, temporally ordered from left to right. Only posts involving at least one of the 10 top-ranked tags are shown. For each tagging event (column), a filled cell marks the presence of the tag in the corresponding row, while an empty cell indicates its absence. A qualitative difference between *blog* (top panel) and *ajax* (bottom panel) is clearly visible, where a higher density at low-rank tags characterizes the semantically narrower *ajax* term. This corresponds to the steeper low-rank behavior observed in the frequency-rank plot for *ajax* (Fig. 3).

Moreover, the low-rank part of the frequency-rank curves exhibits a flattening typically not observed in systems strictly obeying Zipf's law. Several aspects of the underlying complex dynamics may be responsible for this feature: on the one hand this behavior points to the existence of semantically equivalent and possibly competing high-frequency tags (e.g. *blog* and *blogs*). More importantly, this flattening behavior may be ascribed to an underlying hierarchical organization of tags co-occurring with the one we single out: more general tags (semantically speaking) will tend to co-occur with a larger number of other tags. In this scenario, we expect a shallower behavior for tags co-occurring with generic tags (e.g. *blog*) and a steeper behavior for semantically narrow tags (e.g. *ajax*, see also Fig. 2). To better probe the validity of this interpretation, we investigate the co-occurrence relationship that links high-rank tags, lying well within the power-law tail, with low-rank tags located in the shallow part of the distribution. Our observations (see section 5) point in the direction of a non-trivial hierarchical organization emerging out of the collective tagging activity, with each low-rank tag leading its own hierarchy of semantically related higher-rank tags, and all such hierarchies merging into the overall power-law tail.

4. A YULE-SIMON'S MODEL WITH LONG-TERM MEMORY

We now aim at gaining a deeper insight into the phenomenology reported above. In order to model the observed frequency-rank behavior for the full range of ranking values, we introduce a new version of the "rich-get-richer" Yule-Simon's stochastic model [13, 14] by enhancing it with a fat-tailed memory kernel. The original model can be described as the construction of a text from scratch. At each discrete time step one word is appended to the text: with probability p the appended word is a new word, never occurred before, while with probability $1 - p$ one word is copied from the existing text, choosing it with a probability proportional to its current frequency of occurrence. This simple process yields frequency-rank distributions that display a power-law tail with exponent $\alpha = 1 - p$, lower than the exponents we observe in actual data. This happens because the Yule-Simon process has no notion of "aging", i.e. all positions within the text are regarded as identical.

In our construction we moved from the observation that actual users are exposed in principle to all the tags stored in the system (like in the original Yule-Simon model) but the way in which they choose among them, when tagging a new resource, is far from being uniform in time (see also [15, 16]). It seems more realistic to assume that users tend to apply recently added tags more frequently than old ones, according to a memory kernel which might be highly skewed. Indeed, recent findings about human activities [11] support the idea that the access pattern to the past of the system should be fat-tailed, suggesting a power-law memory kernel.

We tested this hypothesis with real data extracted from *del.icio.us*: Fig. 4 shows the temporal auto-correlation function for the sequence of tags co-occurring with *blog*. Such a sequence is constructed by consecutively appending the tags associated with each post, respecting the temporal order of posts. Correlations are computed inside three consecutive windows of length T, starting at different times t_w,

$$C(\Delta t, t_w) = \frac{1}{T - \Delta t} \sum_{t=t_w+1}^{t=t_w+T-\Delta t} \delta(\text{tag}(t + \Delta t), \text{tag}(t)),$$

where $\delta(\text{tag}(t + \Delta t), \text{tag}(t))$ is the usual Kronecker delta function, taking the value 1 when the same tag occurs at times t and $t+\Delta t$. From Fig. 4 it is apparent that the cor-
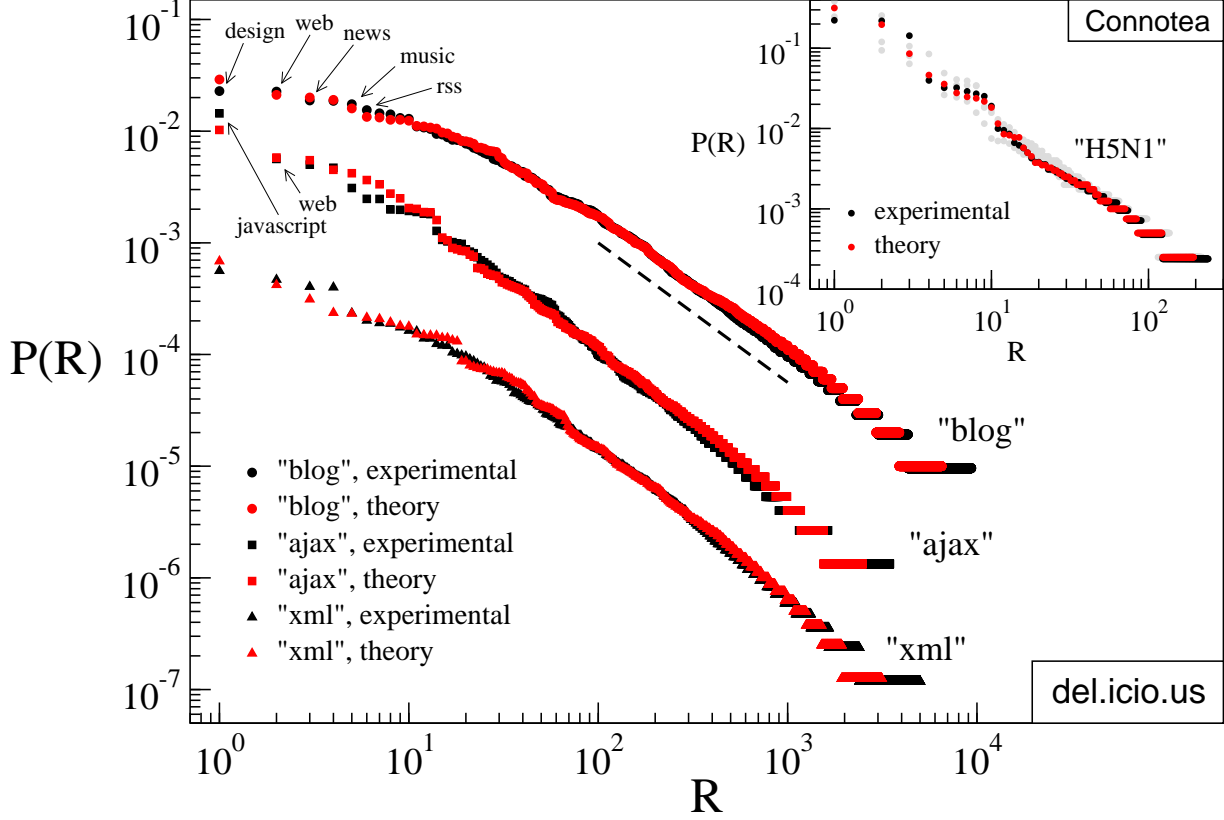
FIG. 3: Frequency-rank plots for tags co-occurring with a selected tag: experimental data (black symbols) are shown for *del.icio.us* (circles for tags co-occurring with the popular tag *blog*, squares for *ajax* and triangles for *xml*) and *Connotea* (inset, black circles for the *H5N1* tag). For the sake of clarity, the curves for *ajax* and *xml* are shifted down by one and two decades, respectively. Details about the experimental datasets are reported in Table I. All curves exhibit a power-law decay for high ranks (a dashed line corresponding to the power law $R^{-5/4}$ is provided as an aid for eye) and a shallower behavior for low ranks. To make contact with Fig. 2, some of the highest-frequency tags co-occurring with *blog* and *ajax* are explicitly indicated with arrows. Red symbols are theoretical data obtained by computer simulation of the stochastic process described in the text (Fig. 5). The parameters of the model, i.e. the probability p, the memory parameter τ and the initial number of words n_0 were adjusted to match the experimental data, giving approximately $p = 0.06$, $\tau = 100$ and $n_0 = 100$ for *blog*, $p = 0.03$, $\tau = 20$ and $n_0 = 50$ for *ajax*, and $p = 0.034$, $\tau = 40$ and $n_0 = 110$ for *xml*. Inset: *Connotea* is a much younger system than *del.icio.us* and the corresponding dataset is smaller and noisier. Nevertheless, a good match with experimental data can be obtained for $p = 0.05$, $\tau = 120$ and $n_0 = 7$ (red circles), demonstrating that our model also applies to the early stages of development of a folksonomy. Gray circles correspond to different realizations of the simulated dynamics.

relation function is non-stationary over time. Moreover, for each value of the initial time t_w a power-law behavior is observed: $C(\Delta t, t_w) = a(t_w)/(\Delta t + \delta(t_w)) + c(t_w)$, where $a(t_w)$ is a time-dependent normalization factor and $\delta(t_w)$ is a phenomenological time scale, slowly increasing with the "age" t_w of the system. $c(t_w)$ is the correlation that one would expect in a random sequence of tags distributed according to the frequency-rank distribution $P_{T,t_w}(R)$ pertaining to the relevant data window. Denoting by $R = R_{\max}(T, t_w)$ the number of distinct tags occurring in the window $[t_w, t_w + T]$, we have $c(t_w) = \sum_{R=1}^{R=R_{\max}(T,t_w)} P_{T,t_w}^2(R)$.

Our modification of the Yule-Simon's model thus consists in weighting the probability of choosing an existing word (tag) according to a power-law kernel. This hypothesis about the functional form of the memory kernel is also supported by findings in Cognitive Psychology [17], where power laws of latency and frequency have been shown to model human memory.

Summarizing, our model of users' behavior can be stated as follows: the process by which users of a collaborative tagging system associate tags to resources can be regarded as the construction of a "text", built one step at a time by adding "words" (i.e. tags) to a text initially comprised of n_0 words. This process is meant to model the behavior of an effective average user in the context identified by a specific tag. At a generic (discrete) time step t, a brand new word may be invented with probability p and appended to the text, while with probability $1 - p$ one word is copied from the existing text, going
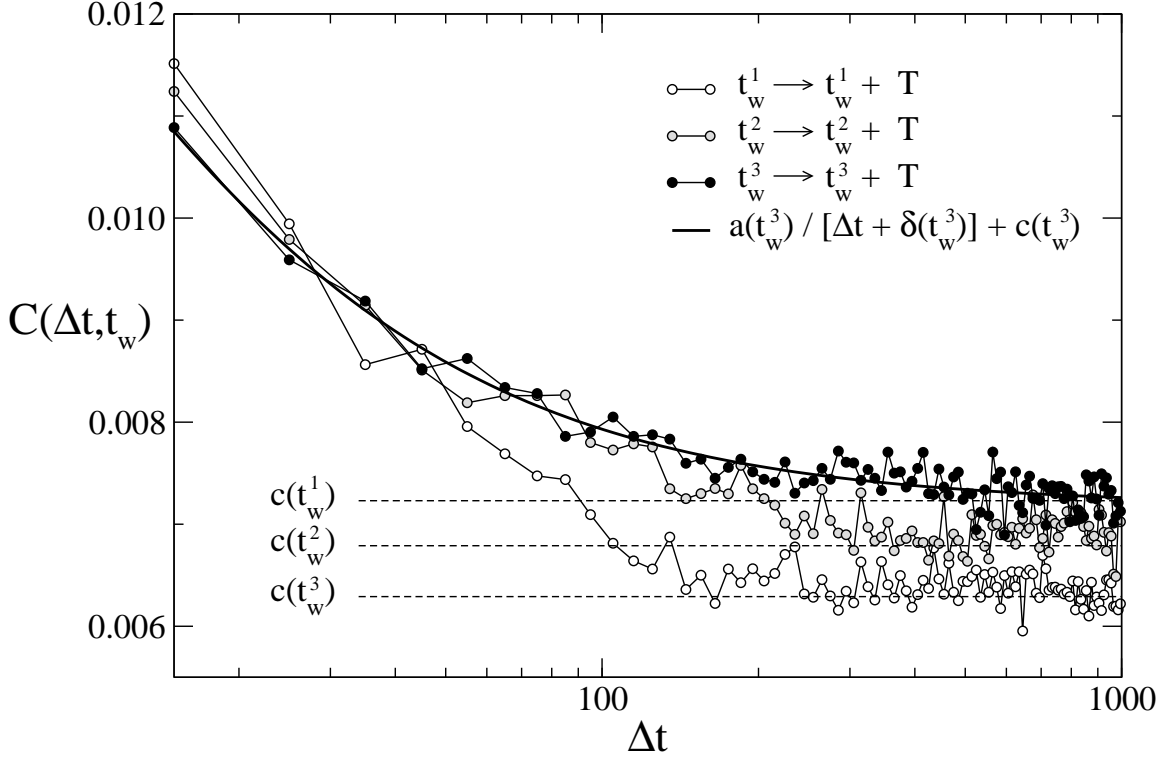
FIG. 4: Tag-tag correlation functions and non-stationarity. The tag-tag correlation function $C(\Delta t, t_w)$ is computed over three consecutive and equally long ($T = 30000$ tags each) subsets of the *blog* dataset, starting respectively at positions $t_w^1 = 10000$, $t_w^2 = 40000$ and $t_w^3 = 70000$ within the collected sequence. Short-range correlations are clearly visible, slowly decaying towards a long-range plateau value. The non-stationary character of correlations is visible both at short range, where the value of the correlation function decays with t_w, and at long range, where the asymptotic correlation increases with t_w. The long-range correlations (dashed lines) can be estimated as the natural correlation present in a random sequence containing a finite number of tags: on using the appropriate ranked distribution of tag frequencies within each window (see text) the values $c(t_w^1)$, $c(t_w^2)$ and $c(t_w^3)$ can be computed, matching the measured plateau of the correlation functions. The thick line is a fit to the fat-tailed memory kernel described in the text.

back in time by x steps with a probability $Q_t(x)$ that decays as a power law, $Q_t(x) = a(t)/(x + \tau)$. $a(t)$ is a normalization factor and τ is a characteristic time scale over which recently added words have comparable probabilities. Fig. 3 shows the excellent agreement between the experimental data and the numerical predictions of our Yule-Simon's model with long-term memory. Our model, unsurprisingly, also reproduces the temporal correlation behavior observed in real data (see [18] for a discussion of this point).

The interpretation of τ (similar to that of the δ parameter introduced above for tag-tag correlations) is related to the number of equivalent top-ranked tags perceived by users as semantically independent (see section 5). In our model, in fact, the average user is exposed to a few roughly equivalent top-ranked tags and this is translated mathematically into a low-rank cutoff of the power law, i.e. the observed low-rank flattening.

Fitting the parameters of the model, in order to match its predictions (obtained by computer simulation) against the experimental data, we obtain an excellent agreement for all the frequency-rank curves we measured, as shown

in Fig. 3. This is a clear indication that the tagging behavior embodied in our simple model captures some key features of the tagging activity. The parameter τ controls the number of top-ranked tags which are allowed to co-occur with comparable frequencies, so that it can be interpreted as a measure of the "semantic breadth" of a tag. This picture is consistent with the fact that the fitted value of τ obtained for *blog* (a rather generic tag) is larger than the one needed for *ajax* (a pretty specific one). Additional information on the role of τ as well as that of p in the framework of our model are reported in [19].
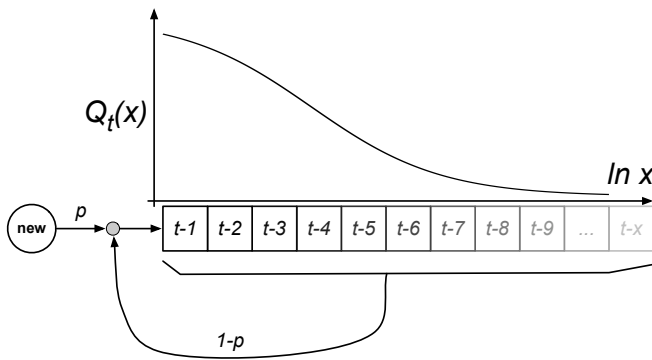
FIG. 5: A Yule-Simon's process with long-term memory. A synthetic stream of tags is generated by iterating the following step: with probability p a new tag is created and appended to the stream, while with probability $1 - p$ a tag is copied from the past of the stream and appended to it. The probability of selecting a tag located x steps into the past is given by the long-range memory kernel $Q_t(x)$, which provides a fat-tailed access to the past of the stream.

5.   CO-OCCURRENCE BETWEEN HIGH-RANK AND LOW-RANK TAGS

Fig. 6 shows a table where the occurrence of 30 high-rank (low-frequency) tags is related to the occurrence of the 15 lowest-rank (highest-frequency) tags. All the tags under study are co-occurring with the tag *blog* and the dataset used for the analysis is the same as the one used in Fig. 2. The co-occurrence analysis is performed as follows: given a high-rank tag X, all resources tagged with X (within the above dataset) are selected, and the co-occurrence frequencies of X with each of the 15 top-ranked (most frequent) tags are recorded. Thus, each row of the table associates a tag X with the corresponding (normalized) co-occurrence histogram. This provides a statistical characterization of tag X in terms of the top-ranked tags, regarded as a natural basis for categorization (or semantic "grounding"). Fig. 7 graphically illustrates such a "tag fingerprint" for 5 high-rank tags, arbitrarily chosen. This analysis is aimed at probing the existence of non-trivial co-occurrence relationships that might be ascribed to semantics and – possibly – to the emergence of a self-organized hierarchy of tags. As shown by the bold numbers in Fig. 6, as well as by the graph in Fig. 7, high-frequency (low-rank) tags do not trivially co-occur with most of the low-frequency (high-rank) tags — on the contrary, the co-occurrence profile of the latter is peaked in correspondence of specific, semantically related tags (*economics* and *law* with *politics*, for example, see Fig. 7). Moreover, several low-frequency (high-rank) tags never co-occur with some of the highest-frequency (low-rank) tags, as shown by the several zeros in Fig. 6. This suggests that high-frequency tags partition – or "categorize" – the resources marked by tags of lower frequency. Given that our definitions of "high-rank" and "low-rank" are somehow arbitrary, and given the self-similar charac-

ter of tag association we observed (Fig. 3), we expect our observations to be representative of a general and complex semiotic structure underlying folksonomies.

6.   CONCLUSIONS

Uncovering the mechanisms governing the emergence of shared categorizations or vocabularies in absence of global coordination is a key problem with significant scientific and technological potential. Collaborative tagging provides a precious opportunity to both analyze the emergence of shared conventions and inspire the design of large (human or artificial) agent systems. Here we report a statistical analysis of tagging activity in a popular social bookmarking system, and introduce a simple stochastic model of user behavior which is able to reproduce the measured co-occurrence properties to a surprisingly level of accuracy. Our results suggest that users of collaborative tagging systems share universal behaviors which, despite the intricacies of personal categorization, tagging procedures and user interactions, appear to follow simple activity patterns. In addition to the findings reported and discussed in this paper, our approach constitutes a starting point upon which studies of greater complexity can be based, with the final goal of understanding, predicting and controlling the Semiotic Dynamics of online social systems.

	design	web	news	music	rss	css	daily	art	politics	tech	technology	blogs	software	media	programming
inspiration	**0.392**	0.159	0.026	0.011	**0.000**	0.206	0.011	0.132	**0.000**	0.011	0.021	0.032	**0.000**	**0.000**	**0.000**
socialsoftware	0.066	0.242	0.022	0.011	0.099	0.044	**0.000**	0.022	**0.000**	0.055	0.055	0.121	0.176	0.077	0.011
economics	**0.000**	0.119	0.090	**0.000**	**0.000**	**0.000**	0.119	**0.000**	**0.552**	0.030	0.045	0.030	**0.000**	0.015	**0.000**
opensource	0.081	0.144	0.108	0.018	0.063	0.009	0.027	0.009	**0.000**	0.054	0.045	0.090	0.243	0.036	0.072
computer	0.167	0.080	0.093	0.013	0.027	0.073	0.033	0.020	0.020	0.080	0.133	0.060	0.127	0.007	0.067
python	0.031	0.138	0.077	**0.000**	0.092	0.046	**0.000**	**0.000**	**0.000**	0.015	0.046	**0.000**	0.092	**0.000**	**0.462**
tagging	0.234	0.213	0.021	**0.000**	0.085	0.085	0.021	**0.000**	0.021	0.021	0.085	0.043	0.170	**0.000**	**0.000**
comics	0.089	0.089	0.089	0.054	**0.000**	0.018	0.054	**0.482**	0.018	**0.000**	0.036	0.054	**0.000**	0.018	**0.000**
research	0.096	0.135	0.096	**0.000**	0.058	**0.000**	0.058	0.038	0.019	0.154	0.077	0.096	0.096	0.096	0.038
law	0.012	0.061	0.134	0.037	**0.000**	**0.000**	0.110	**0.000**	**0.378**	0.024	0.195	0.049	**0.000**	**0.000**	**0.000**
hack	0.080	0.080	0.069	0.034	0.057	0.046	0.126	0.023	0.034	0.161	0.103	0.057	0.069	0.023	0.034
xhtml	**0.311**	**0.262**	0.016	0.004	0.016	**0.336**	**0.000**	0.004	**0.000**	0.008	0.012	0.004	0.012	0.004	0.008
humour	0.125	0.042	0.125	**0.000**	**0.000**	**0.000**	0.125	0.208	0.083	0.083	0.042	**0.000**	**0.000**	0.042	0.125
management	0.147	0.191	**0.000**	**0.000**	0.029	0.015	0.118	0.015	**0.000**	0.015	0.103	0.103	0.162	0.029	0.074
movies	0.080	0.080	0.240	0.120	**0.000**	**0.000**	0.100	0.140	0.080	**0.000**	0.020	0.100	**0.000**	0.040	**0.000**
diy	0.151	0.081	0.047	0.023	**0.000**	**0.000**	0.081	0.186	0.012	0.198	0.093	0.047	0.058	0.012	0.012
life	0.231	**0.000**	**0.000**	0.019	**0.000**	**0.000**	0.096	0.250	0.058	0.058	0.250	**0.000**	0.019	**0.000**	0.019
tag	0.125	**0.354**	**0.000**	**0.000**	0.250	0.021	**0.000**	0.021	**0.000**	0.021	0.062	0.021	0.083	0.021	0.021
maps	0.200	0.150	0.050	**0.000**	0.050	**0.000**	**0.000**	0.050	0.050	0.250	0.050	0.050	0.050	0.050	**0.000**
ideas	0.167	0.103	0.090	**0.000**	0.026	0.077	0.077	0.051	0.038	0.064	0.103	0.077	0.051	0.064	0.013
architecture	**0.526**	0.063	0.074	0.011	**0.000**	**0.000**	0.095	0.137	**0.000**	**0.000**	0.032	0.011	0.011	**0.000**	0.042
organization	**0.000**	0.042	**0.000**	0.021	**0.000**	**0.000**	**0.479**	0.021	**0.000**	0.062	0.125	0.021	0.146	0.042	0.042
plugin	0.156	**0.281**	**0.000**	0.031	0.031	0.031	**0.000**	**0.000**	**0.000**	**0.000**	0.031	0.031	**0.312**	0.031	0.062
blogroll	0.054	0.089	0.107	0.071	0.054	**0.000**	0.054	0.054	0.071	0.089	0.107	0.125	0.054	0.018	0.054
information	0.065	**0.280**	0.172	**0.000**	0.097	0.022	0.022	0.011	0.043	0.054	0.075	0.054	0.054	0.032	0.022
articles	0.159	0.136	0.091	0.011	**0.000**	0.068	0.057	0.011	0.091	0.057	0.080	0.045	0.091	0.045	0.057
resource	**0.287**	0.202	0.032	0.032	0.032	0.191	0.021	0.032	0.011	0.053	0.011	0.011	0.032	**0.000**	0.053
illustration	**0.339**	0.104	0.009	0.026	**0.000**	0.078	0.017	**0.400**	**0.000**	**0.000**	**0.000**	0.017	0.009	**0.000**	**0.000**
mobile	0.021	0.085	0.106	0.021	0.021	**0.000**	0.128	0.106	**0.000**	0.106	**0.255**	0.043	0.064	0.043	**0.000**
liberal	**0.000**	**0.000**	0.125	**0.000**	**0.000**	**0.000**	0.067	**0.000**	**0.692**	**0.000**	**0.000**	0.058	**0.000**	0.058	**0.000**

FIG. 6: Co-occurrence table: columns correspond to the 15 top-ranked tags co-occurring with *blog*, in descending order of frequency from left to right. Rows correspond to 30 low-frequency tags co-occurring with *blog* (frequencies ranking between 100th and 200th). Each row is a normalized co-occurrence histogram representing a "categorization" of the corresponding tag in terms of the top-ranked tags. Numbers in red (bold face) denote co-occurrence probabilities in excess of 25%. Zeros (no co-occurrence) are marked in blue (bold face).
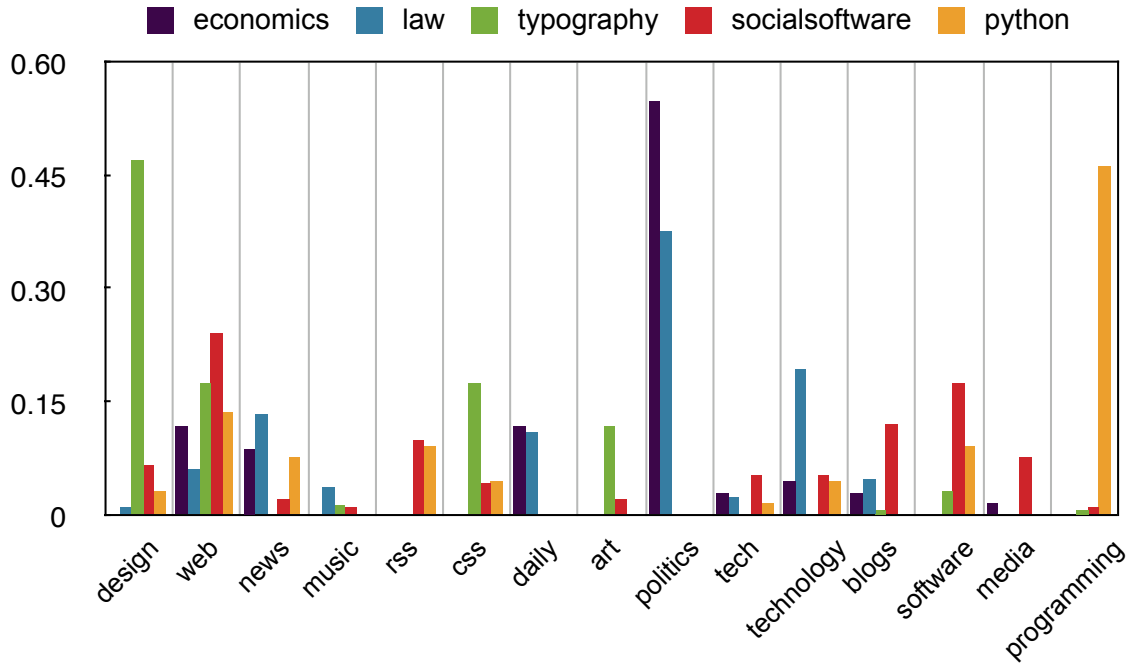

FIG. 7: Co-occurrence patterns for 5 of the low-frequency (high-rank) tags of Fig. 6 (see legend at the top). The colored bars display the "fingerprint" of the selected tags in terms of their co-occurrence with the 15 top-ranked tags (the same ones reported in the top row of Fig. 6).

[1] A. Mates, *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*, Computer Mediated Communication, LIS590CMC.

[2] T. Hammond, T. Hannay, B. Lund and J. Scott, *Social Bookmarking Tools (I): A General Review*, D-Lib Magazine **11**(4), (2005).

[3] S. Golder and B. A. Huberman, *Usage patterns of collaborative tagging systems*, Journal of Information Science **32**(2), 198 (2006).

[4] L. Steels and F. Kaplan, *Collective learning and semiotic dynamics*. In D. Floreano and J.-D. Nicoud and F. Mondada, editors, Advances in Artificial Life: 5th European Conference (ECAL 99), Lecture Notes in Artificial Intelligence **1674**, 679-688, Berlin, (1999).

[5] J. Ke, J. W. Minett, A. Ching-Pong, W.S.-Y, Wang, *Self-organization and selection in the emergence of vocabulary*, Complexity **7**, 41-54 (2002).

[6] M. A. Nowak, N. L. Komarova and P. Niyogy, *Computational and evolutionary aspects of language*, Nature **417**, 611-617 (2002).

[7] S. Kirby, *Natural language and artificial life*, Artificial Life **8**, 182-215 (2002).

[8] T. Vander Wal, *Explaining and Showing Broad and Narrow Folksonomies*, `http://www.personalinfocloud.com/2005/02/explaining_and_.html`, (2005)

[9] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, (1949).

[10] M. E. J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics **46**(5), 323-351 (2005).

[11] A.-L. Barabasi, *The origin of bursts and heavy tails in human dynamics*, Nature **435**, 207 (2005).

[12] R. Ferrer Cancho and V. D. P. Servedio, *Can simple models explain the Zipf's law for all exponents?*, Glottometrics **11**, 1 (2005).

[13] G. U. Yule, *A Mathematical theory of evolution based on the conclusions of Dr. J.C. Willis*, Philos. Trans. R. Soc. London B **213**, 21-87 (1925).

[14] H. A. Simon, *On a class of skew distribution functions*, Biometrika **42**, 425 (1955).

[15] D. H. Zanette and M. A. Montemurro, *Dynamics of text generation with realistic Zipf's distribution*, Journal of Quantitative Linguistics **12**, 29-40 (2005).

[16] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of networks with aging of sites*, Phys. Rev. E **62**, 1842 (2000).

[17] J. R. Anderson, *Cognitive Psychology and its implications*, fifth edition, Worth Publisher, New York (2000).

[18] C. Cattuto and V. Loreto, *Semiotic Dynamics in Online Social Communities*, in preparation (2006).

[19] C. Cattuto, V. Loreto and V. Servedio, *A Yule-Simon process with memory*, submitted (2006).